DOCUMENT RESUME

ED 037 701 AL 002 072

AUTHOR	Bailey, Richard W.; Robinson, Jay L.
TITLE	The Computer in Lexicography.
PUB DATE	24 Oct 69
NOTE	22p.; Paper presented to the section on Computer Research in Language and Literature of the Midwest Modern Language Association, October 24, 1969
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.20
DESCRIPTORS	*Computational Linguistics, *Computer Science, *Data Bases, Dictionaries, Information Processing, *Lexicography, Research Methodology

ABSTRACT

Proposals for the use of the computer in the humanities often ask more of the machine than it can reasonably yield, and the enthusiastic generation of data for dictionary projects may well overburden the editors who must eventually cope with it. Procedures in lexicography are not well enough defined for a substantial burden to be placed on the logical capabilities of the computer. Most data collection must still be left in the hands of human readers, though editing of the data may be carried on with the use of on-line devices in which man interacts with machine. The use of the computer in the final stages of producing a dictionary, however, may yield important results in speeding production and in making available a reservoir of data for other purposes. This paper will be published as part of a "Festschrift" for Hans Kurath later this year. (Author/FB)

ED037701

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The Computer in Lexicography[1]

Richard W. Bailey and Jay L. Robinson

The University of Michigan

Submitted to the editors of a proposed Festschrift for Professor Hans Kurath, February 1970.

AL 002 072

Almost all humanistic scholars who undertake work with the computer claim as a by-product of their research the clarification of fuzzy questions in their discipline. Whether they are concerned with historical, literary, or linguistic matters, they predict that the crooked shall be made straight, the rough places plain. In the last decade a great variety of massive projects have been undertaken in the hope that mechanical data processing techniques can increase the scholar's power to do his proper job and eliminate the meticulous and unproductive sorting or copying of paper slips that occupies so much time in the work of the dialectologist (Shuy 1966) and the lexicographer (Bailey and Robinson 1970). Some of these projects must be regarded as spectacular failures: some have collapsed because of a too sanguine estimate of the extent to which the computer can share the scholar's burden; others through neglect of the careful planning characteristic of the early stages of the Linguistic Atlas of the United States and Canada and of the Middle English Dictionary.

A severe appraisal of recent efforts in humanistic computing will surely not be long in coming, but on this occasion we will pass over the shortcomings of this work. Nevertheless it should be noted that humanistic disciplines already oriented toward statistical or iterative methods have produced the most interesting results so far. At the moment, we have no really successful parsing systems -- much less mechanical translators -- and proposed systems for semantic analysis are very far from turning out descriptions that are of interest to the linguist, the critic, or the lexicographer. The relation between

explicitness and traditional humanistic pursuits is a reciprocal one: objectification and clarification of interesting questions works both ways, both from the computer and from the discipline itself. Any major undertaking -- such as a scholarly dictionary -- must take into account the present state of the art. If the questions to be asked are already well-defined, some success may reasonably be expected. But if the questions are not clearly formulated, a much more cautious view is certainly called for as the humanist approaches the machine.

While a general critical review of computer-based humanistic studies has yet to be carried out, the burgeoning projects in machine translation of the late fifties have been subjected to quite thorough scrutiny. In addition to the intellectual failures previously suggested, many of these projects also failed to justify themselves on economic grounds, at least insofar as the goal of routine mechanical translation of documents was concerned. A committee of the National Academy of Sciences appointed in 1964 declared itself 'puzzled by a rationale for spending substantial sums of money [some $20 million] on the mechanization of a small and already economically depressed industry with a full-time and part-time labor force of less than 5,000' (National Academy of Sciences 1966:12). This committee recognized that many mechanical translation projects made contributions to general linguistic theory and to particular problems of language analysis; therefore it concluded that 'it is wise to press forward undaunted, in the name of science, but that the motive for doing so cannot sensibly be any forseeable improvement in practical translation' (24). Scholarly lexicography, while not obviously a depressed industry, must carefully weigh the

economic justification for the necessarily large expenditures attendant on the use of computers. Both intellectual and economic considerations must assume substantial roles in the lexicographer's thinking as he begins to scrutinize the promise of technological aids to his craft.

Operating under the guise of an apparently scientific method, lexicography is a notoriously ill-defined science.[2] Writing dictionary entries -- the lexicographer's central task -- requires a polymath, sensitive to linguistic nuance and wise in all the lore that people using language talk about. 'No editor,' as Ernst Leisi has recently pointed out, 'has ever produced a theory of definition nor, for that matter, been explicit about the methods that have led him to his conclusions' (Leisi 1964:15). It would be a great mistake, we believe, to think of the computer as a potential 'simulator' of a dictionary editor's behavior. Nonetheless, there are tasks that can be usefully delegated to the machine to take some of the harmful drudgery out of lexicography.

Three major stages can be distinguished in the making of dictionaries: 1) collection of the data on which the dictionary will be based; 2) preparation of the entries, including the choice of canonical forms, the writing of pronunciations, usage notes, definitions, and, in the case of historical and unabridged dictionaries, the selection of illustrative examples from the citation file; and 3) the production of the finished work. What contribution can the computer make to each of these three stages?

1. Establishing a file of usage. Collecting citations for a dictionary has traditionally been carried out by carefully trained readers -- an army of volunteers in the case of the OED and the period and regional descendants of that dictionary, or a substantial group of house readers like that maintained by the G. & C. Merriam Company in Springfield. For reasons that can be easily imagined, a large number of qualified and willing volunteer readers is more difficult to recruit today than it was in the last century, or even a generation ago when the late Professor Charles C. Fries sent out the call for readers for his proposed Early Modern English Dictionary. Varying standards of transcription and accuracy vitiate the usefulness of collections made by independent scholars and much apparently valuable material had to be ultimately abandoned by Professor Kurath in the early stages of the Middle English Dictionary; similar difficulties also vexed Walter S. Avis in his work on the Dictionary of Canadianisms (Avis 1969). Therefore it is no surprise that the possibility of using a computer to take over the laborious work of establishing a file of usage has come to mind wherever new dictionary projects have been initiated.

A thorough survey of those centers where the computer is now at work in establishing a file of citations for dictionary makers would be a considerable undertaking, but we do want to suggest just how widespread this use of the computer is. At Nancy, the Centre de Recherche pour un Trésor de la Langue Française employs thirty-eight full-time clerks in the transcription of texts for their citation file, a collection that will eventually contain excerpts from 250 million

words of text (Centre National de la Recherche Scientifique 1967). A similarly substantial project employing the fullest use of automation is underway in Jerusalem in work for the proposed Historical Dictionary of the Hebrew Language (Ben-Hayyim 1966). The computer likewise plays a prominent role in excerpting texts for a Dictionary of Middle High German (Raben 1969:296), a Historical Dictionary of the Italian Language (Raben 1967:79), and the Dictionary of the Older Scottish Tongue in Edinburgh (Aitken and Pratley 1966). Work is now beginning for a new Old English Dictionary at Toronto and Waterloo, Ontario, that will make significant use of the computer in collecting, and similar lexicological projects are being undertaken for several Quechumaran languages (Wölck 1969), for Latin, Swedish, Dutch, Modern German, Serbo-Croatian, and Russian (see National Science Foundation 1969 and Raben 1969). Nearly all of these projects employ something like the concordance principle to prepare materials for the dictionary editor's hands, and in all cases the work is being undertaken with great vigor -- and undoubtedly great expense.

Despite this widespread activity, we feel that there are some serious limitations to the use of the computer for establishing a citation file and that enthusiasm for the machine has sometimes overburdened the better judgment of the scholar. Professor Louis Milic of Cleveland State, for example, reports that he was asked to become a volunteer reader for a proposed English dictionary of the Tudor period. Incredulous that he was asked to copy citations in longhand from texts in the period, he announces that he found

the editor's request 'baffling.' 'So jarred was I,' he says, 'that I actually began by acquiescing, just to see what sort of work it was. I can report that it was tedious, exacting and repetitive -- just the sort of thing that a computer does very well' (Milic 1967:144). Of course Professor Milic is a pioneer in literary and linguistic computing and he is thoroughly familiar with the powers -- and weaknesses -- of the machine for work of this sort. But our own experience in work on the Early Modern English Dictionary leads us to the conclusion that the computer does this tedious, exacting, and repetitive work all too well. The result of an unbridled automatic reading program for even a small number of texts will simply inundate the editor with material and postpone, rather than hasten, the production of a dictionary.

An excerpt from the directions for OED readers suggests the complexity of a reasonable reading program for a dictionary. Much of the decision-making task is left to that most human of attributes, judgment:

[set in smaller typeface]

Make a quotation for _every_ word that strikes you as rare, obsolete, old-fashioned, new, peculiar, or used in a peculiar way.

Take special note of passages which show or imply that a word is either new and tentative, or needing explanation as obsolete or archaic, and which thus help fix the date of its introduction or disuse.

Make _as many_ quotations _as you can_ for ordinary words, especially when they are used significantly, and tend by the context to explain or suggest their own meaning. (_OED_ I:xv)

It is quite obvious that no mechanical reading program can be devised to extract just those peculiar usages that editors want, or even to do a successful job at gathering 'defining quotations' for normal uses of ordinary vocabulary. Consider, for example, a use of the definite article that apparently arose in the Early Modern period and is illustrated in the OED by a quotation from Nashe: "To borrowe some lesser quarry of elocution from the Latine." How could the computer be sensibly programmed to select this use of *the* with the name of a language without increasing the total number of citations with yards of *the*? Lexicographers prize these special usages and the computer cannot approach the skill of even the most witless volunteer in collecting them unless some sort of oblique strategy is employed in searching for them. Human readers, as the experience of the editors of the Merriam-Webster dictionaries has shown (Macdonald 1962:171-72), are particularly adept at extracting novelties of the kind required by the OED directions. Extracting citations from texts, then, would seem to be a job in which cooperation between man and machine is required; human readers can be asked to identify peculiar usages while the computer can be expected to extract more representative examples of the use of vocabulary. Once a complete concordance has been made, a simple routine can be devised to yield a random selection of frequently occurring words to supplement a collection made by human readers.

Mechanized reading programs, we suggest, can be allowed to bear the whole burden of citation collecting only in those cases where the total corpus for which the dictionary is responsible is quite limited. Old

English is a good example of such a case and the two editors of the new Old English Dictionary can build a citation collection based on virtually the whole body of Old English texts without making their editorial job unmanageable. For more recent periods of our language, a brute force approach in which all usages are isolated is impracticable. The expense of keyboarding more than 10,000 texts for Early Modern English (the OED corpus for the period) can hardly be justified; much less can a lexicographer do more than sample the huge volume of printed language for which a dictionary of a more recent period is responsible.

Nevertheless, the potential for using the computer in building the citation file opens opportunities that were beyond the reach of earlier lexicographers. In our materials for the Early Modern English Dictionary, we have more than two million slips obtained from volunteer readers that we believe to represent a considerable proportion of those special usages requested by the OED editors. Half a million more result from "saturation reading" of some fifty texts selected for their particular linguistic significance, and of course nothing is so useful for saturation reading as the uncomplaining computer. Where scholars like William A. Elwood of Virginia are generous in allowing us to use Renaissance texts keypunched for other purposes, we can produce handsome slips ready for the file quickly and easily. The slips we have prepared using Elwood's version of "The Defense of Poetry" cost about three and a half cents each, according to the current

accounting scheme used by the Computing Center at the University of Michigan. Nearly a decade ago, a commercial lexicographer estimated the cost of gathering slips by traditional means at more than thirty cents each and thanks to economic inflation the present-day cost must be considerably higher (see Barnhart 1962:167). As a result of the programming skills of Dr. Victor J. Streeter, our computer-produced slips match the format of man-made ones, providing on each five by eight slip a generous context for the selected word, full bibliographical information, and a note on the keypunching conventions for special characters. By using a random number generator, we can extract a manageable selection of frequently occurring words rather than excluding them altogether as is usually the case with concordance projects. More extensive use of this computer system will help build a file that will yield valuable information for the editing process without making the collection too enormous for its intended use.

2. Easing the editor's burden. The editing process is intellectually the most interesting part of dictionary making and the most baffling. Though one might hope that the computer could at least provide the editor with a rough sort of the material in the file, in practice the difficulties are enormous. Even the job of separating grammatical homographs like abstract (noun, verb, and adjective) is a difficult one, and the task of separating the senses of polysemous words -- helm of a ship from helm, a helmet -- seems at the moment almost insuperable. In studies now underway, we are exploring the possibility

of gathering textual variants under their canonical form, though without enormously increasing the complexity and expense of the program we cannot expect complete success in this effort. In supposing that the computer can assume the whole burden of grammatical and semantic analysis, we have once again expected too much of the computer and too little of ourselves.

The most promising area of research in computational linguistics, we believe, focuses on the interaction between man and machine. The computer is asked to do what it can do best, to manipulate the linguist's data in useful but not overly complex ways. Essentially the machine rearranges the raw material into word indices, phonetic or lexical concordances, or lists matching word with gloss in an annotated text. No work of any analytical interest is done by the computer, but the data is made more tidy and less forbidding to work with (see Kay 1969). Even more interesting applications are found in the area of hypothesis testing; systems are now in operation that accept proposed syntactic or phonological rules in a format familiar to the linguist, process the rules in a given order, and generate sequences that can be tested for acceptability against the native speaker's intuition (see Friedman et al. forthcoming; Bobrow and Fraser 1968). Both of these applications come close to finding the ground where the best talents of man and machine can operate most effectively.

Lexicographers are only beginning to make use of interactive systems of this sort. At the University of Wisconsin, a system has been devised for the Dictionary of American Regional English that allows the

editor to manipulate the data stored on magnetic tape from a remote terminal, eliminating the time and energy-wasting processes of hand sorting and filing. High costs and minor technical difficulties may yet prevent the use of this system in the actual editing process, but the DARE editors have demonstrated the possibility of using computerized procedures from the optical scanning of prepared data through the final editing process (Venezky 1968). Proofreading the recent American Heritage Dictionary was carried out with the aid of a cathode ray tube terminal which allowed editors to alter the dictionary entries in machine storage. This system came into play only after the main editorial work had been completed, but the use of a terminal of this sort in earlier stages of composition is certainly possible (Publisher's Weekly 1969). The ultimate success of all such systems, however, depends on the ability of thinking lexicographers to rationalize his tasks and his procedures.and to balance economies with results.

In our work at Michigan, we have experimented with an interactive system for editing developed by Dr. Walter Reitman (see Reitman et al. 1969). Citation slips previously encoded appear on a small television screen and the editor writes a provisional definition for that usage. This definition automatically becomes a node on a defining tree and subsequent citations can be assigned to that node or to a new provisional definition written by the editor. When all the citations have been examined, the editor can then arrange the clusters of citations in a way that leads to the heirarchy of senses in the published dictionary. This system is relatively cheap to operate and does not require

extraordinarily expensive hardware. Even less costly systems are promised shortly (Bitzer and Skaperdas 1969), and the regular use of aids of this kind may soon become commonplace in lexicography and in the preparation of scholarly editions.

3. Producing the finished copy. Once the material has been collected and the editorial process completed, the computer may also find a role in the production of the finished dictionary through operation of computer-driven typesetting machines. Commercial lexicographers have been anticipating this development for a decade and only last minute difficulties prevented the production of the Random House Dictionary of the English Language by this means (see Urdang 1966). The recent development of an information system capable of handling typographical complexities made it possible to produce the handsome American Heritage Dictionary by such means, and cheaper systems will soon make the benefits of these schemes available to the scholarly lexicographer as well. The Wycliffe Bible Translators in Mexico City have shown the feasibility of producing testaments in a great variety of special typefaces and styles, and similar systems could be used to handle the complex format that we have come to expect in our English dictionaries.

Economy is not the only benefit to be anticipated from computer production. Standards of textual accuracy can be raised considerably by designing routines to check the editor's finished product; cross-referencing, for example, could be carefully scrutinized, thus saving the editor from one source of animadversion by sharp-eyed reviewers.

Should the final stages of dictionary-making be mechanized, it would not be difficult to produce abridgements of various sorts to fill the demands of students as well as scholars. But even more important, the computer-encoded dictionary would allow regular revision and alteration as new evidence comes to light. As a by-product of lexicography, the tape or disc containing the dictionary might eventually play an important role in an enormous question-answering system of the kind now envisaged by many scholars in information science. Though these techniques and applications are not of great interest to the lexicographer, they do suggest that lexicography, as well as other applications of humanistic computing, may eventually come to play an important part in the text-editing schemes, translators, and information systems of the future.

Other forms of modern technology may soon have an impact on the lexicographer's work. Scholarly dictionaries are notoriously slow projects; of the six dictionaries that were begun in response to Craigie's call for regional and period dictionaries in 1919, only one is now completed (the Dictionary of American English) and more than a quarter century of work finds the Dictionary of the Older Scottish Tongue, the Scottish National Dictionary, and the Middle English Dictionary still far from completion. The scholarship that makes use of these works cannot stand still and some interim means is needed for disseminating the lexical information now locked in their files. Microform technology offers the possibility of making the

citation files for these dictionaries available at major research centers, not only as an aid to literary scholars but also for use by lexicographers interested in other periods of the language. Furthermore, the availability of vocabulary citations on microfilm would offer the scholar interested in particular usages more information than a selection from the file in a published dictionary could possibly give him. The cost and size of a scholarly dictionary might also be reduced by a system of reference keys between the dictionary and the microfilm collection, an innovation that might well hasten the production of such works.

Rising costs and new styles of scholarship have combined to inhibit the initiation of new lexicographical projects. Yet recent years have seen the production of two new historical dictionaries for English, the Dictionary of Canadianisms and the Dictionary of Jamaican English. Similar undertakings are afoot in Australia and New Zealand, and A. W. Read shortly expects to begin editing his Dictionary of Briticisms, a project first announced in 1938. R. W. Burchfield's supplementary volume to the OED is now in production, while other projects have been tentatively outlined (see Crystal 1967 and Országh 1967). Technology derived from the computer and from micromethods can help speed the completion of such important accounts of regional and period English. The impact of technology will be greater, however, and more significant if British and American dictionary makers will follow the example of their continental colleagues, as Sledd suggests,

in treating technological innovation as merely a catalyst for a general examination of traditional practices. The computer is only a tool, but like any tool it can be used to overcome inherent deficiencies and extend limited powers.

[footnotes]

[1]An abridged version of this paper was presented to the section on Computer Research in Language and Literature of the Midwest Modern Language Association, October 24, 1969.

[2]Some linguists have recently turned to the problems of lexicography and its relation to contemporary theories of language. In addition to well-known attempts to specify the nature of the lexicon in a generative grammar, works by the following authors (listed at the end of this essay) are of interest in considering the practical problems of dictionary making from a linguistic point of view: Hiorth, Hoffer, Householder and Saporta, Lebrun, Meier, Országh, Pottier, and Tollenaere.

REFERENCES

Aitken, A. J., and Paul Pratley. 1966. An Archive of Older Scottish Texts for Scanning by Computer. Studies in Scottish Literature 4.45-47.

Avis, Walter S. 1969. Problems Met in Editing the Dictionary of Canadianisms. Unpublished paper presented to the Present-Day English Section of the Modern Language Association.

Bailey, Richard W., and Jay L. Robinson. 1970. Computers and Dictionaries. Computers and Old English Concordances, ed. by Angus Cameron, Roberta Frank, and John Leyerle, 000-00. Toronto, Univ. of Toronto Press.

Barnhart, C. L. 1962. Problems in Editing Commercial Monolingual Dictionaries. Problems in Lexicography, ed. by Fred W. Householder and Sol Saporta, 161-81. Bloomington, Indiana Univ. Research Center in Anthropology, Folklore, and Linguistics, publ. 21.

Ben-Hayyim, Ze'ev. 1966. A Hebrew Dictionary on Historical Principles. Ariel 13.14-20.

Bobrow, Daniel G., and J. Bruce Fraser. 1968. A Phonological Rule Tester. Communications of the Association for Computing Machinery 11.766-72.

Centre National de la Recherche Scientifique. 1967. Centre de Recherche pour un Trésor de la Langue Française. Paris, Bureau des Relations Extérieures et de l'Information.

Crystal, David. 1967. Report of the Pilot Survey for the Proposed 'Dictionary of the English-Speaking Peoples.' Mimeographed.

Friedman, Joyce, et al. forthcoming. A Computer Model of Transformational Grammar. New York, American Elsevier.

Hiorth, Finngeir. 1955a. Arrangement of Meanings in Lexicography. Lingua 4.413-24.

_____. 1955b. On the Subject Matter of Lexicography. Studia Linguistica 9.57-65.

_____. 1956. On the Relation between Field Research and Lexicography. Studia Linguistica 10.57-66.

_____. 1957. On the Foundations of Lexicography. Studia Linguistica 11.8-27.

_____. 1958. On Defining 'Word'. Studia Linguistica 12.1-26.

_____. 1960. Zur Ordnung des Wortschatzes. Studia Linguistica 14.65-84.

Hoffer, Bates L. 1966a. Lexology and Lexicography. GENGO RONSHUU (Tokyo University of Education) 7.89-102.

_____. 1966b. Semology and Webster's Third. EIGO SEINEN 112.74-77, 159-61.

Householder, Fred W., and Sol Saporta, eds. 1962. Problems in Lexicography. Bloomington, Indiana Univ. Research Center in Anthropology, Folklore, and Linguistics, publ. 21.

Kay, Martin. 1969. The Computer System to Aid the Linguistic Field Worker. Santa Monica, RAND paper P-4095.

Lebrun, Yvan. 1963. Lexicographie et structuralisme. Revue Belge de Philologie et d'Histoire 41.815-19.

Leisi, Ernst, ed. 1964. Measure for Measure: An Old-Spelling and Old-Meaning Edition. Heidelberg, Winter.

Macdonald, Dwight. 1962. The String Untuned. Repr. in Dictionaries and THAT Dictionary, ed. by James Sledd and Wilma R. Ebbitt, 166-88. Chicago, Scott, Foresman.

Milic, Louis T. 1967. Making Haste Slowly in Literary Computation. Computers in Humanistic Research, ed. by Edmund A. Bowles, 143-52. Englewood Cliffs, Prentice-Hall.

Meier, Hans Heinrich. 1969. Lexicography as Applied Linguistics. English Studies 50.141-51.

National Academy of Sciences. 1966. Language and Machines: Computers in Translation and Linguistics. Washington, D. C., National Research Council Publication 1416.

National Science Foundation. 1969. Current Research and Development in Scientific Documentation, 15.91-162. Washington, D. C.

Országh, László. 1960. Problems and Principles of the New Dictionary of the Hungarian Language. Acta Linguistica Academiae Scientiarum Hungaricae 10.211-73.

_____. 1967. A Plea for a Dictionary of Modern Idiomatic English. Hungarian Studies in English 3.71-81.

Pottier, Bernard. 1965. "La définition sémantique dans les dictionnaires. Travaux de Linguistique et de Littérature 3.33-39.

Publisher's Weekly. Sept. 1, 1969. Producing a $4 Million Book. 56-58.

Raben, Joseph, compiler. 1967. Directory of Scholars Active. Computers and the Humanities 2.71-93.

_____. 1969. Directory of Scholars Active. Computers and the Humanities 4.125-41.

Read, Allen Walker. 1938. Plans for 'A Historical Dictionary of Briticisms'. The American Oxonian 25.186-90.

Shuy, Roger W. 1966. An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada. Computation in Linguistics, ed. by Paul L. Garvin and Bernard Spolsky, 60-75. Bloomington, Indiana Univ. Press.

Sledd, James. 1968. Toward the First New International Alvearie. Unpublished paper presented to the Present-Day English section of the Modern Language Association.

Tollenaere, F., de. 1963. Nieuwe Wegen in de Lexicologie. Amsterdam, Noord-Hollandsche Uitgevers Maatschappij.

Urdang, Laurence. 1966. The Systems Designs and Devices Used to Process the Random House Dictionary of the English Language. Computers and the Humanities 1.31-33.

Venezky, Richard L. 1968. Storage, Retrieval, and Editing of Information for a Dictionary. American Documentation 19.71-79.

Wölck, Wolfgang. 1969. A Computerized Dictionary of Andean Languages. Language Sciences no. 8.1-9.